J. Pijar MIPA, Vol. V No.2, September : 81 - 85
ISSN 1907-1744

PENDEKATAN ESTIMATOR KERNEL UNTUK ESTIMASI DENSITAS MULUS

Laila Hayati
Program Studi Pendidikan Matematika PMIPA FKIP Universitas Mataram
Jl. Majapahit No. 62 Mataram 83125
e-mail: lailaanugerah@yahoo.com

Abstrak : Misalkan diberikan data pengamatan independen $\{X_i : i = 1,2,..., n\}$ dengan fungsi densitas *f*. Ada dua pendekatan dalam mengestimasi *f* yaitu dengan pendekatan parametrik dan pendekatan nonparametik. Pada pendekatan nonparametrik dilakukan jika asumsi bentuk f tidak diketahui. Dalam hal ini diasumsikan bahwa fungsi *f* termuat dalam kelas fungsi mulus. Salah satu teknik untuk mengestimasi fungsi mulus adalah teknik pemulus kernel. Tingkat kemulusan fungsi estimasi ditentukan oleh parameter pemulus. Semakin besar parameter pemulusnya semakin mulus fungsi estimasi dan sebaliknya.

Kata-kata Kunci: *densitas mulus, estimator kernel, parameter pemulus*

KERNEL ESTIMATOR APPROACHES FOR ESTIMATES SMOOTH DENSITY

Abstract: Let $\{X_i : i = 1,2,..., n\}$ be independent observation data from a distribution with density function *f*. There are two basic approaches for estimating *f*, the parametric and the nonparametric approaches. In nonparametric approaches, an unknown density function *f*. The function f is assumed to be a smooth function, so the function f could be estimated by kernel estimator. The smoothing level of kernel estimator depends to the smoothing parameter. The big smoothing parameter gives the estimation function which over smooth and the contrary.

Key Words: *smooth density, kernel estimator, smoothing parameter*

I. PENDAHULUAN

Dalam analisa regresi, tidak semua variabel penjelas dapat didekati dengan pendekatan parametrik, karena tidak adanya informasi yang jelas bagaimana bentuk hubungan variabel penjelas dengan variabel responnya sehingga harus digunakan pendekatan nonparametrik. Tujuan analisa regresi adalah menentukan hampiran untuk kurva regresi *m.*

Jika diberikan data pengamatan independen $\{X_i : i = 1,2,..., n\}$, untuk menentukan distribusi dari X ekivalen dengan menentukan fungsi densitasnya. Untuk mengestimasi fungsi densitas *f* dapat dilakukan dengan dua pendekatan yaitu pendekatan parametrik dan nonparametrik. Pendekatan nonparametrik dilakukan jika asumsi bentuk *f* tidak diketahui. Dalam hal ini diasumsikan bahwa fungsi *f* termuat dalam kelas fungsi mulus yaitu mempunyai turunan kontinu atau terintegralkan secara kuadrat.

Permasalahan dalam densitas nonparametrik adalah bagaimana mengkonstruksikan estimasi dari fungsi densitas tanpa membuat asumsi struktural seperti tentang bentuk fungsi, tetapi hanya mensyaratkan bahwa fungsi densitas tersebut sekurang-kurangnya mempunyai dua turunan yang terbatas.

Salah satu teknik untuk mengestimasi fungsi mulus adalah teknik pemulus kernel [1]. Metode yang paling sederhana adalah histogram. Teknik pemulus kernel pada estimator densitas merupakan pengembangan dari estimator histogram.

Dari Hayati [4], dengan menggunakan pendekatan regresi nonparametrik untuk menemukan estimator untuk fungsi regresi *m* diperoleh estimator yang konsisten (dengan menggunakan kernel normal, *bandwidth* 0.01; 0.1; dan 1), yaitu estimasi total populasi semakin mendekati total populasi dengan semakin bertambahnya jumlah sampel (*n*). Semakin besar *bandwidth*, maka estimasi total populasi semakin menjauhi total populasi.

Dalam tulisan ini dibahas tentang pencarian estimator kernel dari densitas mulus, sifat-sifat dan contoh simulasinya dengan program S-Plus for Windows.

Estimator Histogram [2]

Metode estimasi densitas secara nonparametrik yang paling populer adalah histogram. Namun sebenarnya, histogram ini bukanlah merupakan alat estimasi densitas yang baik, karena bentuknya yang sangat mudah dipengaruhi oleh jumlah kelas dan lokasi nilai tengahnya, dan juga estimasi densitas yang dihasilkan tidak kontinu pada batas kelas. Diketahui sampel random $\{X_i : i = 1,2,..., n\}$ dari suatu populasi dengan fungsi densitas tak diketahui f. Berdasarkan sampel random ini akan diestimasi fungsi densitasnya. Misalkan daerah nilai *x* dibagi menjadi disjoin interval-interval dengan panjang *2h*. Peluang observasi yang masuk ke dalam interval $(X_0 - h, X_0 + h)$ adalah:

$$P\{X \in (X_0 - h, X_0 + h)\} = \int_{X_0 - h}^{X_0 + h} f(x)\,dx$$

Diperoleh estimator histogram untuk *f(x)* yaitu:

$$\hat{f}_h(x) = \frac{1}{2nh} \#\{X_i \in (X_0 - h, X_0 + h)\}$$

untuk semua $X \in (X_0 - h, X_0 + h)$. ini berarti bahwa observasi yang masuk ke dalam interval yang tergantung *h*, yakni $(X_0 - h, X_0 + h)$ memberi sokongan yang sama besar terhadap $f_h(x)$. pemilihan lebar kelas *h* kecil, histogram memuat banyak batang kecil-kecil, sedangkan untuk *h* besar histogram memuat sedikit batang besar-besar.

Estimator Kernel [2]

Fungsi kernel *K* yang umum dipakai adalah fungsi densitas dan biasanya dilengkapi dengan asumsi-asumsi tertentu. Jika $X = (X_1, X_2, ..., X_n)$ sampel random dari suatu distribusi densitas *f* dan K suatu fungsi terbatas dan positif yang memenuhi sifat sebagai berikut:

$$\int_{-\infty}^{\infty} y_i K(y)dy = \begin{cases} 1, \text{ jika } i = 0 \\ 0, \text{ jika } 1 \le i < r-1, \text{untuk suatu bilangan } r \\ \neq 0, \text{ jika } i = r \end{cases}$$

Maka fungsi *K* yang memenuhi sifat di atas disebut dengan Kernel berorder-*r*. sifat-sifat lainnya adalah bahwa *K* merupakan fungsi densitas dan simetrik sekitar nol. Ini didasarkan atas kenyataan bahwa:

$\int_{-\infty}^{\infty} K(y)dy = 1$ dan $\int_{-\infty}^{\infty} yK(y)dy = 0.$ Kernel ini akan digunakan untuk mengkonstruksikan estimator densitas nonparametrik dari $f_h(x)$, yaitu:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) ...(*)$$

Dengan *h* penghalus kernel, yang akan memegang peranan dalam mendefinisikan estimator $\hat{f}_h(x)$ dan menentukan variansi dan biasnya.

Beberapa contoh fungsi Kernel:

1. Kernel Uniform: $K(u) = \frac{1}{2} I(|u| \le 1)$
2. Kernel Triangle: $K(u) = (1 - |u|) I(|u| \le 1)$
3. Kernel Epanechnikov: $K(u) = \frac{3}{4}(1 - u^2) I(|u| \le 1)$
4. Kernel Quartic: $K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \le 1)$
5. Kernel Triweight: $K(u) = \frac{35}{32}(1 - u^2)^3 I(|u| \le 1)$
6. Kernel Gaussian:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) I_{(-\infty,\infty)}(u)$$

dimana *I* adalah fungsi karakteristik.

$$I(|u| \le 1) = \begin{cases} 1, & jika \quad |u| \le 1 \\ 0, & jika \quad |u| > 1 \end{cases}$$

Contoh:

Kernel Uniform: $K(u) = \frac{1}{2} I(|u| \le 1)$

$$K(u) = \begin{cases} \frac{1}{2}, & jika \quad |u| \le 1 \\ 0, & jika \quad |u| > 1 \end{cases}$$

akan ditunjukkan bahwa $\int K(u)du = 1$

$$\int_{-\infty}^{\infty} K(u)du = \int_{-\infty}^{-1} 0du + \int_{-1}^{1} \frac{1}{2}du + \int_{1}^{\infty} 0du = 1$$

Sifat-sifat Statistik Densitas Kernel

Misalkan $\{X_i\}_{i=1}^{N}$ pengamatan variabel random yang berdistribusi independen dan identik, dengan densitas *f*. Estimasi densitas kernel berdasarkan dua parameter yaitu:

- Bandwidth *h*
- Fungsi densitas kernel *K*

Dalam estimator kernel, parameter penghalus *h* merupakan pengontrol keseimbangan antara kesesuaian kurva terhadap data dan kemulusan kurva, maka sangat penting untuk menentukan h_{opt} sehingga estimator yang diperoleh juga optimal.

Berikut diuraikan sifat-sifat statistik densitas kernel.

Teorema 1.1 [3]

Jika $\hat{f}_h(x)$ *diberikan oleh persamaan (*), maka untuk* $h \to 0$, $\hat{f}_h(x)$ *tak bias secara asimtotis.*

Bukti:

Karena X_i berdistribusi independen dan identik maka:

$$\begin{aligned} E(\hat{f}_h(x)) &= E\left(\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(K_h(x - X_i)) \\ &= E(K_h(x - X)) \\ &= \int_{-\infty}^{\infty} K_h(x - u) f(u)du \\ &= \int_{-\infty}^{\infty} K(s) f(x + sh)ds, \end{aligned}$$

dengan substitusi $u = x + sh$,

untuk $h \to 0$, diperoleh :

$$E(\hat{f}_h(x)) \to \int_{-\infty}^{\infty} K(s) f(x)ds = f(x)$$

Jadi estimasinya tak bias secara asimtotis.

Sifat bias dapat juga dianalisis menggunakan ekspansi Taylor dari $f(x + sh)$ disekitar *x*, yang diasumsikan $f \in C^2$ (kontinu diferensiabel dua kali).

Teorema 1.2 [3]

Jika $\hat{f}_h(x)$ *diberikan oleh persamaan (*), maka* $Bias(\hat{f}_h(x)) = \frac{h^2}{2} f''(x)\mu_2(K) + o(h^2), h \to 0$ *dengan* $\mu_2(K) = \int s^2 K(s)ds$.

Bukti:

$$\begin{aligned} Bias(\hat{f}_h(x)) &= \int_{-\infty}^{\infty} K(s) f(x + sh)ds - f(x) \\ &= \int_{-\infty}^{\infty} K(s)\left[f(x) + \frac{hs}{1!}f'(x) + \frac{h^2 s^2}{2!}f''(x) + o(h^2)\right] ds - f(x) \\ &= f(x)\int_{-\infty}^{\infty} K(s)ds + hf'(x)\int_{-\infty}^{\infty} sK(s)ds + \frac{h^2}{2}f''(x) \\ &\quad \int_{-\infty}^{\infty} s^2 K(s)ds + o(h^2) - f(x) \\ &= f(x) + \frac{h^2}{2}f''(x)\mu_2(K) + o(h^2) - f(x) \end{aligned}$$

dimana,

$$o(h^2)=\frac{\frac{(sh)^3}{3!}f'''(\xi)}{h^2}\to 0, untuk h\to 0, dengan\ x<\xi<x+sh$$

Jadi estimasi densitas kernel:

$$Bias\left(\hat{f}_h(x)\right)=\frac{h^2}{2}f''(x)\mu_2(K)+o(h^2), h\to 0$$

Terlihat bahwa bias merupakan fungsi kuadrat dalam *h*. Oleh karena itu, dibutuhkan *h* yang kecil untuk menurunkan biasnya.

Teorema 1.3 [3]

Jika $\hat{f}_h(x)$ *diberikan oleh persamaan (*), maka*

$$\text{var}\left(\hat{f}_h(x)\right)=(nh)^{-1}\|K\|_2^2 f(x)+o\left((nh)^{-1}\right), nh\to\infty$$

dengan $\|K\|_2^2=\int K^2(s)ds$.

Bukti:

Karena X_i berdistribusi independen dan identik maka:

$$Var\left(\hat{f}_h(x)\right)=Var\left(\frac{1}{n}\sum_{i=1}^{n}K_h(x-X_i)\right)$$

$$=\frac{1}{n^2}\sum_{i=1}^{n}\left(\text{var}\ K_h(x-X_i)\right)$$

$$=\frac{1}{nh^2}\left[E\left(K^2\left(\frac{x-X}{h}\right)\right)-E^2\left(K\left(\frac{x-X}{h}\right)\right)\right]$$

$$=\frac{1}{nh^2}\left[\int K^2\left(\frac{x-u}{h}\right)f(u)du-\left(\int K\left(\frac{x-u}{h}\right)f(u)du\right)^2\right]$$

$$=\frac{1}{n}\left[\frac{1}{h}\int K^2(s)f(x+sh)ds-\left(\int K(s)f(x+sh)ds\right)^2\right]$$

$$=\frac{1}{n}\left[\frac{1}{h}\int K^2(s)f(x+sh)ds-(f(x)+o(h))^2\right]$$

$$=\frac{1}{n}\left[\frac{1}{h}\|K\|_2^2(f(x)+o(h))-(f(x)+o(h))^2\right]$$

$$=\frac{1}{nh}\|K\|_2^2 f(x)+o\left((nh)^{-1}\right),\ nh\to\infty$$

dimana,

$$o(h)=\frac{\frac{(sh)^2}{2!}f''(\xi)}{h}\to 0, untuk\ h\to 0,\ dengan\ x<\xi<x+sh$$

$$o\left((nh)^{-1}\right)=\frac{(nh)^{-1}\|K\|_2^2 o(h)-\frac{h(f(x)+o(h))^2}{nh}}{(nh)^{-1}}\to 0,$$

$$untuk\quad (nh)^{-1}\to 0,\ nh\to\infty$$

Terlihat bahwa variansi proporsional dengan $(nh)^{-1}$. Oleh karena itu, dibutuhkan *h* yang besar untuk menurunkan variansinya. Hal ini bertentangan dengan biasnya, sehingga diperhatikan *MSE* yang memberikan kontrol antara bias kuadrat dan variansi.

II. PEMBAHASAN

Telah diketahui secara umum, bahwa permasalahan utama pada pemulus kernel tidak erletak pada pemilihan kernel tetapi pada pemilihan *bandwidth*. Pemilihan *bandwidth* optimum lebih ditekankan pada penyeimbangan antara bias dan varians. Satu perumusan masalah yang dapat memperlihatkan hubungan antara bias dan varians adalah *MSE*, karena itu dengan meminimumkan *MSE* maka masalah antara bias dan varians dapat diminimumkan juga.

Teorema 2.1 [3]

Jika $\hat{f}_h(x)$ *diberikan oleh persamaan (2.6), maka*

$$MSE\left(\hat{f}_h(x)\right)=\frac{1}{nh}\|K\|_2^2 f(x)+\frac{h^4}{4}(f''(x)\mu_2(K))^2+o\left((nh)^{-1}\right)+o\left(h^4\right),\ h\to 0,\ nh\to\infty$$

Bukti:

Dari persamaan *(2.3),* maka $MSE\left(\hat{f}_h(x)\right)=Var\left(\hat{f}_h(x)\right)+Bias^2\left(\hat{f}_h(x)\right)$. Dengan menggunakan teorema *2.6* dan *2.7* maka diperoleh:

$$MSE\left(\hat{f}_h(x)\right)=\frac{1}{nh}\|K\|_2^2 f(x)+\frac{h^4}{4}(f''(x)\mu_2(K))^2+o\left((nh)^{-1}\right)+o\left(h^4\right),\ h\to 0,\ nh\to\infty$$

Teorema 2.2 [3]

Jika $h\to 0,\ nh\to\infty$, *maka* $\hat{f}_h(x)$ *adalah estimator konsisten untuk f(x).*

Bukti:

Dari teorema *3.1,* terlihat bahwa jika $h\to 0,\ nh\to\infty$, maka $MSE\left(\hat{f}_h(x)\right)\xrightarrow{p}0$. Dengan kata lain $\hat{f}_h(x)\xrightarrow{p}f(x)$. Selanjutnya didefinisikan bandwidth optimal h_{opt}, yang diperoleh dari,

$$h_{opt}=\arg\min_h MSE\left(\hat{m}_h(x)\right)$$

sehingga diperoleh teorema berikut:

Teorema 2.3 [3]

Jika $h\to 0, nh\to\infty$, *maka:*

(i). $h_{opt}=o\left(n^{-1/5}\right)$

(ii). $MSE\left(\hat{m}_{hopt}\right)=o\left(n^{-4/5}\right)$

Bukti:

(i). Dari teorema **[3],** dikatakan bahwa

Jika , maka Pendekatan MSE untuk $\hat{m}_h(x)$ *adalah*

$$MSE(\hat{m}_h(x))=\frac{1}{nh}\frac{\sigma^2(x)}{f(x)}\|K\|_2^2+\frac{h^4}{4}\left(m''(x)+2\frac{m'(x)f'(x)}{f(x)}\right)^2\mu_2^2(K)+o\left(nh^{-1}\right)+o\left(h^4\right)$$

diperoleh:

$$MSE(\hat{m}_h(x)) = \frac{1}{nh}A + \frac{h^4}{4}B$$

dengan,

$$A = \frac{\sigma^2(x)}{f(x)}\|K\|_2^2$$

$$B = \left(m''(x) + 2\frac{m'(x)f'(x)}{f(x)}\right)^2 \mu_2^2(K)$$

$$\frac{\partial MSE\ (\hat{m}_h(x))}{\partial h} \approx \frac{1}{nh^2}A + h^3 B$$

Apabila diambil $\frac{\partial MSE\ (\hat{m}_h(x))}{\partial h} = 0$, diperoleh:

$$h_{opt} = \left(\frac{A}{nB}\right)^{1/5}$$

$$= \left(\frac{\sigma^2\|K\|_2^2}{f(x)n\left(\left(m''(x) + \frac{2m'(x)f'(x)}{f(x)}\right)^2 \mu_2^2(K)\right)}\right)^{1/5}$$

$$= o\left(n^{-1/5}\right)$$

(ii). Apabila nilai h_{opt} disubstitusi ke $MSE\ (\hat{m}_{h\,opt})$ diperoleh:

$$MSE\ (\hat{m}_{h\,opt}) = \left(\frac{A}{n}\right)^{4/5} B^{1/5} + \frac{1}{4}\left(\frac{A}{n}\right)^{4/5} B^{1/5}$$

$$= \frac{5}{4}\left(\frac{\sigma^2}{f(x)n}\|K\|_2^2\right)^{4/5} \left(\left(m''(x) + \frac{2m'(x)f'(x)}{f(x)}\right)^2 \mu_2^2(K)\right)^{1/5}$$

$$= o\left(n^{-4/5}\right)$$

Contoh simulasi estimasi densitas Kernel

Bagaimana pengaruh fungsi Kernel (Triangle, Parzen, Gaussian) dan *bandwidth (h=0.01; 0.05; 0.1; 0.5).*

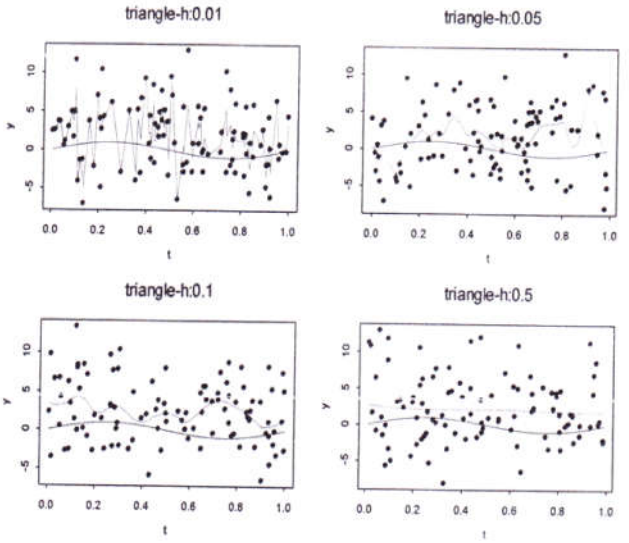
Misalkan: $Y = \sin(2\pi t)$

$t = sort(unif(0,1))$

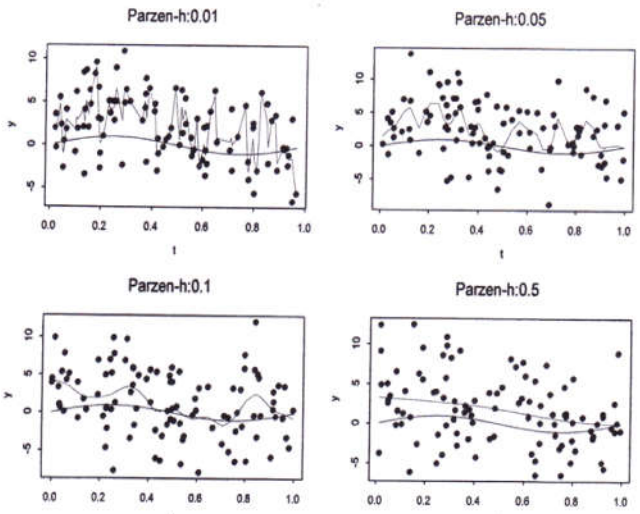
$\varepsilon \sim N(2,4)$

Estimasi densitas nonparametrik dengan menggunakan kernel di atas ditunjukkan pada gambar berikut ini:

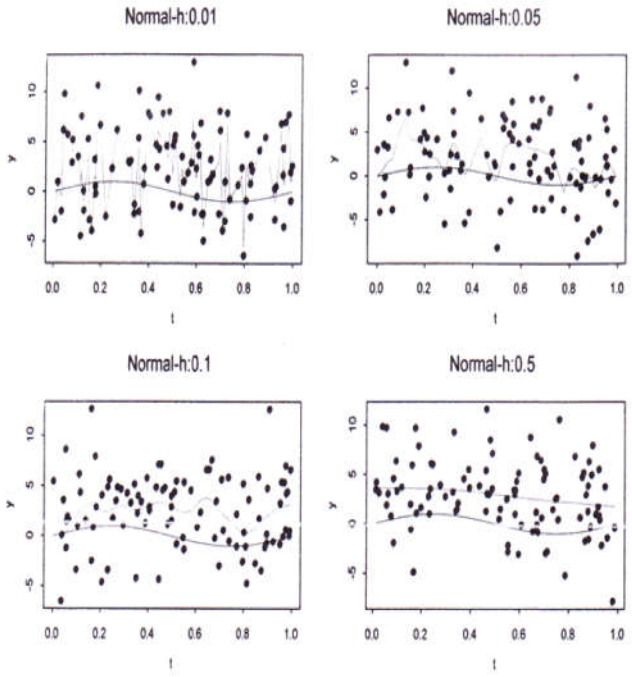
Gambar 1. Estimasi Densitas dengan Kernel Triangle



Gambar 2. Estimasi Densitas dengan Kernel Parzen



Gambar 3. Estimasi Densitas dengan Kernel Gaussian



Dari tampilan gambar 1, 2, dan 3 dapat dibuat kesimpulan yaitu:

1. Dengan *bandwidth* yang tetap, tidak ada perbedaan signifikan secara visual (grafik) dengan berubahnya fungsi kernel.
2. Dengan fungsi kernel yang tetap, terdapat perbedaan yang signifikan secara visual (grafik) dengan berubahnya *bandwidth*. Dimana semakin besar bandwidth maka garis grafik yang dihasilkan semakin mulus.

III. KESIMPULAN DAN SARAN

Dari uraian di atas, dapat disimpulkan bahwa untuk mengestimasi fungsi densitas *f*, jika informasi model distribusi *X* tak diketahui maka *f* dapat diestimasi dengan menggunakan pendekatan nonparametrik. Salah satu pendekatan nonparametrik dengan menggunakan teknik pemulus kernel. Tingkat kemulusan fungsi estimasi ditentukan oleh parameter pemulus. Semakin besar parameter pemulusnya semakin mulus fungsi estimasinya dan sebaliknya.

Adapun saran yang dapat dikemukakan adalah: perlu dilanjutkan pembahasan pada masalah fungsi kernel yang lainnya selain Kernel Triangle, Kernel Parzen, dan Kernel Gaussian, dengan *bandwidth* yang bervariasi, dan perlu dikaji juga teknik untuk mengestimasi fungsi mulus yang lainnya.

DAFTAR PUSTAKA

[1] Hardle, W. 1990. *Applied Nonparametric Regression Analysis*. Cambridge University Press, Cambridge.

[2] Hardle, W. 1990. *Smoothing T echniques Wih Implementation in S*. Springer Verlag, New York

[3] Hardle, W. 1991. *Sampling Technique* . Springer Verlag, London.

[4] Hayati, L. 2010. Regresi Nonparametrik Untuk Mengestimasi Total Populasi Berhingga. *Jurnal Penelitian Universitas Mataram*. 2 (15): 1-8.